Nazir, S., Yousaf, M.H., Nebel, J.C., Velastin, S.A. (2018). A bag of expression framework for improved human action recognition. *Pattern Recognition Letters*, 103 pp. 39-45.

DOI: 10.1016/j.patrec.2017.12.024

A Bag of Expression Framework for Improved Human Action Recognition

Saima Nazir[a], Muhammad Haroon Yousaf[a], Jean-Christophe Nebel[b], Sergio A. Velastin[c,d,**]

[a] *University of Engineering and Technology, Taxila, Pakistan*
[b] *Kingston University, London, UK*
[c] *Universidad Carlos III de Madrid, Spain*
[d] *Queen Mary University of London, UK*

ABSTRACT

The Bag of Words (BoW) approach has been widely used for human action recognition in recent state-of-the-art methods. In this paper, we introduce what we call a Bag of Expression (BoE) framework, based on the bag of words method, for recognizing human action in simple and realistic scenarios. The proposed approach includes space time neighborhood information in addition to visual words. The main focus is to enhance the existing strengths of the BoW approach like view independence, scale invariance and occlusion handling. BOE includes independent pairs of neighbors for building expressions, therefore it is tolerant to occlusion and capable of handling view independence up to some extent in realistic scenarios. Our main contribution includes learning a class specific visual words extraction approach for establishing a relationship between these extracted visual words in both space and time dimension. Finally, we have carried out a set of experiments to optimize different parameters and compare its performance with recent state-of-the-art-methods. Our approach outperforms existing Bag of Words based approaches, when evaluated using the same performance evaluation methods. We tested our approach on four publicly available datasets for human action recognition i.e. UCF-Sports, KTH, UCF11 and UCF50 and achieve significant results i.e. 97.3%, 99.5%, 96.7% and 93.42% respectively in terms of average accuracy.

1. Introduction

Recently, human action recognition has become an emerging research topic in the computer vision field. The recognition and classification of human actions can play an important role in video surveillance, scene recognition, human computer interaction, video indexing and retrieval etc. An effective action recognition approach is essential for attaining high recognition accuracy. However, it is still a challenging task to recognize human actions in uncontrolled environments subject to occlusion, background clutter, scale variance, and view independence challenges.

The Bag of Words (BoW) approach has shown state-of-the-art performance in video representation for human action recognition. Using local space time features along with the BoW approach has proved advantageous for handling realistic dataset challenges (Wang et al., 2009; Niebles et al., 2008; Wang et al., 2016). These methods represent videos in term of local features instead of tracking any specific object, therefore they can be robust to scale invariance, occlusion and viewpoint variations (Wang et al., 2009). Usually spatio-temporal representations are obtained using different methods including various detectors and descriptors such as 3D Harris (Laptev and Lindeberg, 2003), Cuboid detector (Dollár et al., 2005), Hessian detector (Willems et al., 2008), 3D SIFT (Scovanner et al., 2007), HOG3D (Klaser et al., 2008) and ESURF (Willems et al., 2008) followed by video representation such as histogram of feature occurrence frequency using the BoW method. BoW requires selecting specific parameters for a sampling strategy i.e. extracting localized features from a video sample, size of the code book, quantization, distance function used for in nearest-neighborhood assignment and classifier. In such video representation approaches, recognition performance is highly dependent on the discrimination power of the chosen features (Peng et al., 2016).

**Corresponding author:
e-mail: sergio.velastin@ieee.org (Sergio A. Velastin)

BoW contains information of single visual words but ignores the spatial contextual information of words. Thus, such video representation is not able to express contextual relationships between words, thus limiting overall action recognition accuracy. To tackle this limitation, we propose a novel approach to represent words in groups of words that we call *expressions*, and integrate the spatio-temporal relationship between words. The main idea is to preserve the existing strength of classical BoW and include spatial and temporal information that is usually lost during word formation, by encoding neighborhoods amongst words.

(Kovashka and Grauman, 2010) included neighborhoods amongst words in a classical BoW and obtained state-of-art results in datasets such as KTH (Schuldt et al., 2004) and UCF-Sports (Rodriguez et al., 2008). In their work, they have preserved the scale invariance capability of the classical BoW, but there is little capability of handling different viewpoints and occlusions. Another promising work which uses neighborhood amongst features was proposed by Gilbert et al. (2011). To preserve some scale invariance, the inclusion of neighborhoods requires only the relative position and scale of compound features to form a spatio-temporal hierarchy. Gilbert et al. (2011) employed localized neighborhood grouping of features at the initial stage and the volume of neighborhood was increased in each hierarchy level to obtain compound features. They explicitly addressed the problem of scale invariance by using the relative position and scale of these compound features. In their paper, they represent each detected interest point by a three digit string encoding (scale, channel and orientation) and establish neighborhood by encoding relative position of corners in a 3X3X3 grid. This limits their techniques ability to handle variation in viewpoints.

Peng and Schmid (2016) proposed a multi-region two stream CNN (Convolutional Neural Network) scheme based on three recent methods i.e. two stream CNN, R-CNN and optical flow stacking and multi-region CNNs. For UCF-Sports dataset, they improved recognition performance using detection aware features and precise action localization. In Abdulmunem et al. (2016), they extracted features only from those video frames with salient regions and further described them using 3D SIFT and histogram of oriented optical flow (HOOF) descriptors.

Wang et al. (2013) captured local motion information using dense trajectories and described these extracted features using HOG, HOF and point coordinates to characterize appearance, motion and shape information respectively. The proposed motion boundary histogram descriptors are robust to camera motion. They tested their approach for KTH, Hollywood2, UCF Sports, YouTube, Olympic Sports and UIUC action recognition datasets. The results show that the proposed dense trajectories outperform other trajectory extraction approaches with KLT tracker or SIFT descriptor matching. However, their approach is computationally expensive and its performance is limited by the available optical flow quality. Yang et al. (2015) represented features using multi-scale oriented neighborhood features (MONFs), which are formed by concatenating Single-scale Oriented Neighborhood Features (SONF). They evaluated their approach on KTH and UCF Sports dataset and stated that the proposed method outperforms the local ST features based method for action recognition. Wu et al. (2011) represented videos using two types of distribution features, i.e. multiple Global GMM distribution using relative coordinates between interest points in local regions and GMM distribution of local video appearance. They proposed an augmented feature multiple kernel learning algorithm (AFMKL) for classification purposes and tested their approach on UCF-Sports, Multi-view IXMAS and KTH datasets. Results prove the effectiveness of AFMKL through the use of pre-learned classifiers from other action classes. Generally, multiple kernel learing (MKL) methods assume that training and testing data are from the same feature distribution and domain (Duan et al., 2012). Therefore, training data from an auxiliary domain can degrade the performance of MKL algorithms in the target domain. Finally, Yadav et al. (2016) recognized action by detecting interest points that captures differential motion information. They show the discriminative ability of the proposed approach for different action classes by analyzing the distinctness factor of descriptors. For UCF11, they used temporal localization information for interest point detection and showed the importance of temporal localization for video representation. Further improvement is needed to make this approach robust to scale and view invariance.

In this paper, we propose a new approach, Bag of Expression (BoE), to represent spatio-temporal contextual relationships between words. To improve the handling of view independence along with scale invariance and occlusion challenges present in realistic videos, we pair each visual word with a number of neighbors in the spatio-temporal domain to obtain independent visual word pairs. This strategy also aims at providing tolerance to occlusion as each pair of neighborhood is independent. Inclusion of neighborhood information with words significantly outperforms state of-the-art results for KTH, UCF-Sports, UCF11 and UCF50 datasets. We demonstrate (please see section 3) that our Bag of Expression approach leads to better action recognition in both controlled and realistic environments.

The rest of paper is organized as follows. In section 2 we describe our proposed Bag of Expression approach. We introduce the formation of expressions using neighborhood information between words in addition to learning class specific dictionary. In section 3 we discuss experiments and results including comparisons of our approach with recent state-of-the-art methods.

2. Bag of Expression

The flowchart of our proposed algorithm is shown in fig.1. For training purposes, the process starts by extracting features from labeled videos with unique action class labels as $L=\{L_1,L_2,.....,L_l\}$, where l is the total number of action classes. In the next step, we define visual words to generate a visual expression codebook using the space time neighborhood relationship between extracted visual words. Occurrences of each expression are counted to form histogram of expressions. Finally, BoE based representation of training videos is used to train a supervised classifier.

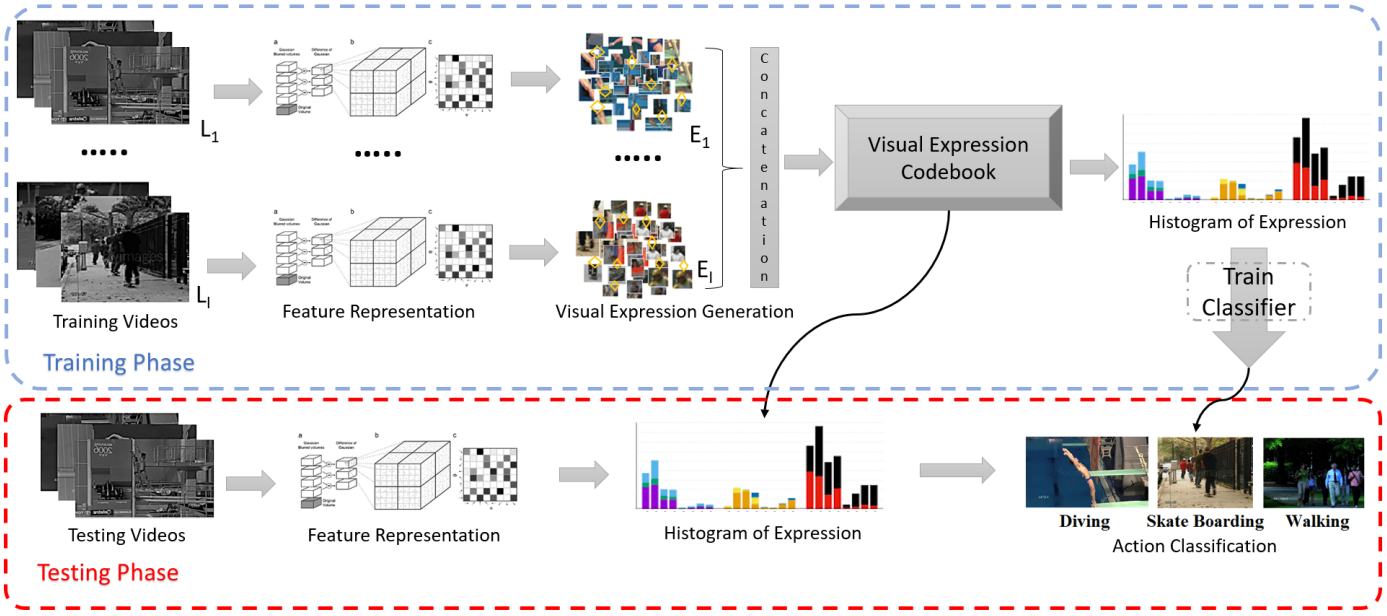

Fig. 1. Novel Bag of Expression framework for human action recognition.

Similarly, during the testing phase, feature representation is obtained for unlabeled videos and quantized using the visual expression codebook created during training. A histogram of expression is created for the occurrence of every expression for each testing video which is then passed to the trained classifier to find the action label. These processes are explained in more detail in the following sub-sections.

2.1. Feature Extraction

We employ a 3D Harris Interest Point detector (Laptev and Lindeberg, 2003), an extension of the Harris detector in the temporal domain, for detection of space time interest point. The 3D Harris Interest point detector has been adopted by many researchers to extract sparse local features. We use this detector to obtain interest points that are well localized in the spatio-temporal domain and corresponds to meaningful events. As shown in fig.2 (for graphical representation simplicity interest points are shown in the space domain only), spatial interest points are detected with a distinct location and have large variation in both space and time dimensions. These interest points correspond to non-constant motion in spatio-temporal neighborhood. (Laptev and Lindeberg, 2003) considered a spatio-temporal second moment matrix, which is a 3x3 dimension matrix of first order spatio-temporal derivatives, along with a weighted Gaussian function $\mathcal{G}(:,{\sigma_i}^2,{\tau_i}^2)$ as previously used in (Nagel and Gehrke, 1998) for optical flow computation. The matrix $\mathbb{M}$ is defined as:

$$\mathbb{M} = \mathcal{G}(:, \sigma_i^2, \tau_i^2) * \begin{Bmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{Bmatrix} \quad (1)$$

Thus, space time interest points are extracted by detecting points which have large eigenvalues in the matrix. Note that in our method we have not adapted detected interest points to scale and velocity in order to obtain sufficient number of interest points so that events can be differentiated from each other and noise, as in (Laptev and Lindeberg, 2003).

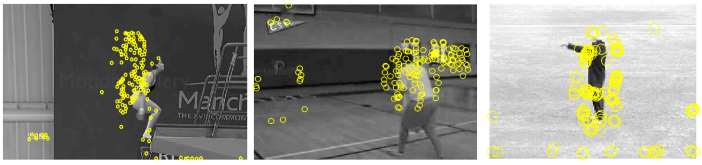

Fig. 2. Detected space time Interest Points on sample videos for UCF Sports (Diving), UCF11 (Basketball) and KTH (Boxing) dataset.

2.2. Feature Description

Once space-time Interest Points are extracted, they are described using the 3-Dimensional Scale Invariant Feature Transform (3D SIFT). 3D SIFT, proposed by Scovanner et al. (2007), is an extension of SIFT (Lowe, 2004) in 3-dimensions, where, as it is common in video processing (Gilbert et al., 2011; Kovashka and Grauman, 2010), the third dimension is time. The presence of occlusions, noise and dynamic background complicates the recognition of actions captured in realistic scenarios. To handle these challenges 3D SIFT provides robustness to noise and orientation by encoding information in both space and time domains. To describe an interest point, 3D gradient magnitude and orientation of its neighborhood are computed, followed by encoding 3D SIFT representation using sub-histograms. 3D gradient magnitude and orientation is represented as:

$$m_3D = \sqrt{(L_x^2 + L_y^2 + L_t^2)} \quad (2)$$

$$\theta(x, y, t) = tan^{-1}\left(\frac{L_y}{L_x}\right) \quad (3)$$

$$\phi(x,y,t) = tan^{-1}\left(\frac{L_t}{\sqrt{L_x^2 + L_y^2}}\right) \quad (4)$$

θ(x,y,t) and ϕ(x,y,t) are used to represent 3D gradient orientations for each pixel. θ(x,y,t) represents angle in 2D gradient and ϕ(x,y,t) represents the angle away from the 2D gradient directions. Orientation information is accumulated into sub-histograms, which are created by sampling surrounding sub-regions of each interest point. The final 3D SIFT representation is obtained by vectorization of these sub-histograms.

2.3. Visual Expression Generation

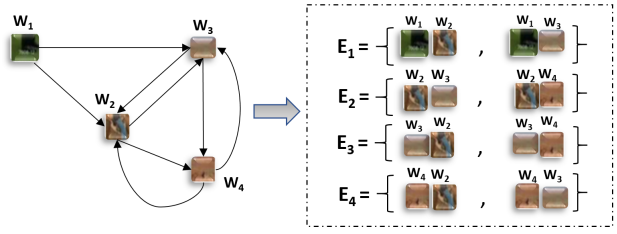

Fig. 3. Visual expression generation (for clarity we have set N=2 and considered only 4 visual words).

After feature representation using 3D Harris and 3D SIFT, each video is represented as V={$(x_1,y_1,t_1,\ p_1)$,$(x_2,y_2,t_2,\ p_2)$,.....,$(x_m,y_m,t_m,\ p_m)$} where (x_h,y_h,t_h) is the spatio-temporal position vector and p_h is the 3D SIFT feature representation of the h^{th} detected interest point and m interest points identified in video V.

We intend to learn a class specific clustering and extract visual words that are most discriminative for a given action class. We represent a set of obtained features as F={F_1,F_2,.....,F_l} , where l is the number of unique action classes and F_i={F_{i1},F_{i2},.....,F_{in_i}}, where n_i is the number of features obtained for the ith action class.

We then apply the widely used k-means clustering (Jain, 2010) on the obtained set of features to divide the feature set F into k clusters. Each cluster center C_q is associated with a visual word and is denoted as w_q.

The visual word set W is composed of the visual words from the l action classes and is represented as W={W_1,...,W_i,.....,W_l},where there are k visual words in the set W_i of each class i (W={W_i=W_{i1},W_{i2},.....,W_{ik}}).

For each visual word representation w_{ij}, we calculate its N nearest neighbors in space time domain by calculating a given distance. We have considered four different distance measures for distance calculation: Mahalanobis, Euclidean, Hamming and City block distances. Our experiments show that the Mahalanobis distance measure shows better performance w.r.t the other three distance measures. For all visual words $w \in W$, the Mahalanobis distance is calculated as:

$$D_M(W_{ij}) = \sqrt{(w, W_{ij})^T S^{-1} (w, W_{ij})} \quad (5)$$

where S is the covariance matrix of the two independent visual words w and W_{ij}. We describe a neighborhood by independent pairs of neighbors. Moreover, these selected nearest neighbors are paired with the respective visual word independently of their relations with the other words. For example, if there are 10 neighbors for a visual word, each neighbor is paired with the visual word to obtain ten different descriptors. This enables some degree of view independence and provides better tolerance to occlusion as each pair of neighbor is an independent word containing spatio-temporal features.

To support view independence, we represent every action using expression and only the frequency of these expressions is stored. Furthermore, it is used to represent the distribution of these expressions in the space time domain. This approach resembles that of the Histogramming methods, which is a simple form to achieve view independence (Weinland et al., 2011). A Histogramming based approach is also used in (Zelnik-Manor and Irani, 2001) to represent instance distribution in space time gradient.

Similarly, representing in the form of expression, independent pair of words, also enables some degree of tolerance to occlusion, as such representation discards all information related to other words and only considers the relation between respective words in a pair, it focuses on the individual contribution of expression and enhances its discrimination power in occluded environment.

As shown in fig.3 each word W_i in set W is paired with its N neighbors and is represented as expression E_i where e_{ij}=W_iW_j.

$$\begin{aligned} E_1 &= \{e_{11}, e_{12}, .., e_{1N}\} \\ E_1 &= \{e_{21}, e_{22}, .., e_{2N}\} \\ &... \\ E_T &= \{e_{T1}, e_{T2}.., e_{TN}\} \end{aligned} \quad (6)$$

Fig. 4. Visual expression and STIPs representation (in space domain only) on the example videos frames of diving, golfswing and kicking actions from UCF Sports dataset.

These expressions are further grouped together and represented as a visual expression code book E={E_1,E_2,.....,E_T}, where E_i={e_{i1},e_{i2},.....,e_{iN}} and T is the total number of expressions. Some visual expression examples are shown in fig.4 for the diving, golf swing and kicking action classes from the UCF-Sports dataset. In this figure, only the information of space domain is mapped on sample frames for graphical representation. As shown in fig.4, expression samples are different with respect to each action classes and are discriminative enough to differentiate one action class representation from another.

Fig. 4(a) shows the location of STIPs for three videos in each action class (i.e. diving, golfswing, kicking). These STIPs locations are shown in the space domain only. Similarly, fig.4 (b) also shows the location of visual expression in the space

domain only, which is the general representation of 'Diving', 'GolfSwing' and 'Kicking' action classes. It might be the case that visual expression samples seem random (fig. 4(b)) however, the visual expression representation is not only dependent on its location (x,y,t) but also on the description of that location (i.e. feature description using 3D SIFT descriptor). Experimental results suggest that even if two different actions occur in the same environment, then the visual expression would not be same as our visual expressions are the representation of actions and not the environment. As shown in fig.5, the spatio-temporal interest points are used to represent actions. Although the presence of unwanted actions in the background can affect visual expressions, results demonstrate they improve performance in comparison with other methods.

Fig. 5. Lifting and Diving actions in same environment(STIPs are represented using red dots).

2.4. Histogram of Expression

Each feature vector f_i is mapped to the nearest visual expression by calculating the Euclidean distance between an expression codebook E and the respective feature vector f_i. The histogram of expression is formed by calculating the occurrence frequency of each expression for video V. Each video V is represented as a histogram of expression and denoted as EV_i={HE_1,HE_2,.....,HE_T}, where HE_i={He_{i1},He_{i2},.....,He_{iN}}. A histogram of expression is shown in fig.6 for UCF Sports dataset for a few action classes. As discussed earlier, class specific words are generated to form expressions. Therefore, each histogram exhibits some class specific properties. For example, for the diving class, there are expressions with high frequency which discriminate its representation from other classes.

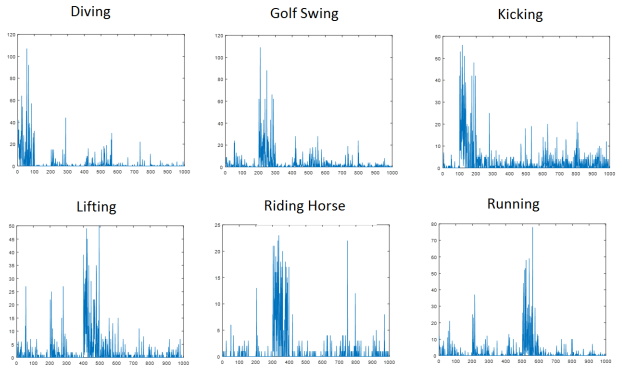

Fig. 6. Histogram of Expression for UCF sports action dataset.

2.5. Action Classification

For classification purposes, we employed a popular non-linear multiclass support vector machine classifier for action classification. This method learns *l(l-1))/2* binary support vector machine (SVM), where l is the number of unique action classes by using one-versus-one coding design model. It improves classification accuracy for multi-class classification by minimizing the aggregation of losses for P binary learner.

$$\widehat{\mathrm{l}} = \arg\min_{l} \frac{|m_{lp}| \sum_{p=1}^{p} g(m_{lp}, s_p)}{\sum_{p=1}^{p} |m_{lp}|} \tag{7}$$

where m is the one-versus-one coding design matrix with element m_{lp}, $\mathrm{s_p}$ is the predicted classification score for the positive class for learner p and $g()$ is the binary loss function.

3. Experimental Results

To determine the effectiveness of our method, we evaluate our approach on four publicly available datasets i.e. UCF50 (Reddy and Shah, 2013), UCF11 (Liu et al., 2009), UCF-Sports (Rodriguez et al., 2008) and the KTH (Schuldt et al., 2004) datasets. For interest point detection and description, we use default parameters settings provided by Laptev and Lindeberg (2003) and (Scovanner et al., 2007).

KTH is a standard dataset for recognizing actions in simple scenarios. It is captured in a controlled environment with simple background and holds camera motion and zooming effect in few videos (Mukherjee et al., 2011). It has 6 action classes i.e. walking, jogging, running, hand waving, hand clapping and boxing. These actions are performed by 25 actors in 4 different environments. We used standard training and testing split as described in (Schuldt et al., 2004) and used by most state-of-the-art methods. Video sequences were divided based on the subjects, 16 subjects video sequences were used for training while the remaining 9 subjects video sequences were used for testing.

UCF-Sports is captured in a more realistic environment than KTH dataset since it contains 150 video sequences taken from real broadcast sport events. UCF-Sports comprises 10 action classes including weight-lifting, swimming, horse riding and golf-swing. It is captured with cluttered backgrounds, different viewpoints, occlusions, motions and scale discontinuities. We use a Leave One Out cross validation method for evaluation purpose following Rodriguez et al. (2008).

UCF11, previously known as YouTube action dataset, and UCF50 (an extension of UCF11 dataset) are captured in realistic environments with large variations in viewpoints, backgrounds, camera motions, object appearances and poses. UCF11 and UCF50 contain 11 and 50 action categories respectively. For each action class, video clips are grouped into 25 groups each containing at least 4 video clips. Each group shares some similar features, like similar environment, same actor and similar viewpoints. As proposed by Rodriguez et al. (2008) and Reddy and Shah (2013), a Leave One Group Out evaluation method is used for evaluation purposes for both datasets.

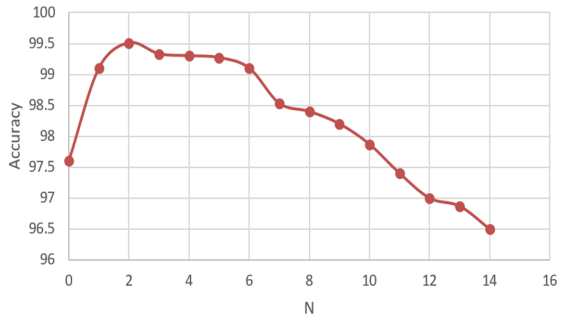

Fig. 7. Average accuracy w.r.t number of neighbors *N* for visual expression codebook formation for KTH dataset using the Mahalanobis distance.

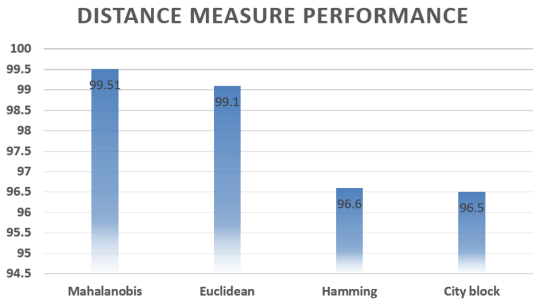

Fig. 8. Performance evaluation using different distance measures for neighbor calculation for the KTH dataset.

3.1. Performance Evaluation

Optimization of parameters involves two different parameters: the number of clusters *k* and the number of nearest neighbors *N* for creation of the expression codebook. We evaluate our approach by varying *k* and then *N* for the KTH dataset. As shown in Table 1 accuracy increases with respect to an increase in the number of clusters *K* up to *k=300*, whereas *N* is constant, i.e. *N=2*. Then, accuracy plateaus while execution time carries on increasing.

For optimization of *N*, we performed different experiments with *k=300* on the KTH dataset. As illustrated in fig. 7, construction of a large number of neighbors leads to creation of non-relevant neighbors, which may contain erroneous information resulting in a loss of performance. While values of *N* between 1 and 6 achieve accuracies above 99%, the best performance is produced for *N=2*. Note that the case *N=0* corresponds to performance of the standard bag of words approach. Its lower accuracy demonstrates the added value of the usage of visual expressions.

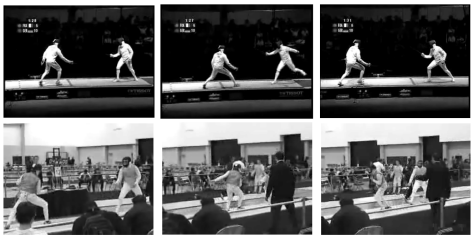

Fig. 9. Positive results for fencing action (UCF50 dataset) in non-occluded and occluded environment

Table 1. Average accuracy performance and execution time w.r.t to different number of cluster (k) for KTH dataset.

K	Accuracy	Time(Secs)
50	94.99	90.5
100	96.22	195.88
200	98.78	362.12
300	99.51	525.47
400	99.51	703.65
500	99.51	870.39
600	99.51	1020.99

Fig.8 shows performance results by using different distance measures for calculation of *N* nearest neighbors in space time domain for the KTH dataset. We evaluate the performance of our proposed approach using four different distance measures, i.e. Mahalanobis, Euclidean, Hamming and City block distances with visual expression codebook size *T=N*k*l=2*300*6=3600* for KTH dataset. As shown in fig.10, the Mahalanobis distance measure shows better results with 99.51% average accuracy for the KTH dataset.

The proposed method also shows some degree of tolerance to occlusion in realistic environments. Fig 9 shows examples of positive results in an occluded environment for the fencing action from the UCF50 dataset. As proposed, BoE only considers the spatio-temporal neighborhood of each visual word by discarding the information related to other visual words representation, therefore focusing on the individual contribution of each expression and enhancing its discrimination power in an occluded environment. As a result, each visual expression is a representation of the local information of an independent patch making it relatively invariant to occlusion

Fig. 10 shows the performance of our method, with *k=300* and *N=2*, on the UCF50, UCF11, UCF Sports and KTH datasets in terms of confusion matrix. It shows that our approach achieved reasonable performance on most of the action classes for both simple and realistic datasets. The decrease in performance in realistic scenarios is expected since three datasets, i.e. UCF Sports, UCF11 and UCF50, contains some unwanted actions in the background, which can mislead classifiers. For the KTH dataset, there is only confusion between walking and jogging class which is intuitive because of the inter class similarity between both actions.

3.2. Comparison with state-of-the-art

We conclude the experimentation and results discussion with comparison of our approach with state-of-the-art methods. As shown in Table 2, our method outperforms other mentioned methods in terms of average accuracy for UCF Sports, KTH, UCF11 and UCF50 datasets. It should be noted that BoE performance was optimized for KTH dataset as discussed in section 3.1. BoE outperforms MultiScale Neighborhood features (MONFs) based approach for UCF Sports and KTH dataset (Yang et al., 2015). MONFs was formed by concatenating Single scale neighborhood features (SONF). Improved results were obtained for all mentioned datasets (i.e. KTH, UCF Sports and UCF11) as compared to BOVW when compared with results

Table 2. Comparison with state-of-the-arts methods for UCF-Sports, KTH, UCF11 and UCF50 Datasets.

Dataset	Paper	Method	Results
UCF Sports	Our	Bag of Expression (BoE)	**97.33**%
	Peng and Schmid (2016)	Multi Region two stream R-CNN	95.74%
	Abdulmunem et al. (2016)	Bag of Visual words	90.90%
	Wang et al. (2013)	Dense Trajectories and motion boundary descriptor	88.00%
	Yang et al. (2015)	Multi-scale oriented neighborhood features	91.80%
	Kovashka and Grauman (2010)	Hierarchical Space time neighborhood features	87.27%
KTH	Our	Bag of Expression (BoE)	**99.51**%
	Abdulmunem et al. (2016)	Bag of Visual words	97.20%
	Wang et al. (2013)	Dense Trajectories and motion boundary descriptor	95.00%
	Gilbert et al. (2011)	Mined Hierarchical compound features	94.50%
	Yang et al. (2015)	Multi-scale oriented neighborhood features	96.50%
	Kovashka and Grauman (2010)	Hierarchical Space time neighborhood features	94.53%
UCF11	Our	Bag of Expression (BoE)	**96.68**%
	Wang et al. (2011)	Dense Trajectories	84.20%
	Yadav et al. (2016)	Motion Boundaries and Dense Trajectories	91.30%
	Mota et al. (2013)	Tenser Motion Descriptor	75.40%
	Liu et al. (2009)	Bag of visual words	71.20%
UCF50	Our	Bag of Expression (BoE)	**93.42**%
	Duta et al. (2017)	HMG + iDT Descriptor	93.00%
	Peng et al. (2016)	Bag of Words and Fusion Methods	92.30%
	Wang et al. (2016)	Dense Trajectories	91.70%
	Wang et al. (2013)	Dense Trajectories and motion boundary descriptor	91.20%

presented in (Abdulmunem et al., 2016) and (Liu et al., 2009). Performance is improved by around 2% for UCF Sports and KTH datasets and, for UCF11 and UCF50, we significantly improve the results by around 5% and 0.42% respectively.

4. Conclusion

Bag of words has proved to be a promising model for real word action recognition problems and it is preferred by many authors due to its simplicity and lack of any requirement for preprocessing input data. In this paper, we have proposed an extension of BOW, calling it Bag of Expression, which includes neighborhood relationship information between words in space and time domain. It utilizes the existing strengths of BOW and learns class specific clustering algorithm by learning neighborhood information that is most discriminative for given action class. BOE enables some degree of view independence and provides tolerance to occlusion as it describes neighborhoods through independent pairs of neighbors containing local spatio-temporal information. We demonstrated the capabilities of our approach for action classification in both simple and realistic scenarios and have shown that BOE outperforms recent state-of-the-art methods. For future work we will explore the use of deep learning methods to learn the impact of space time neighborhood information for efficient action recognition.

Acknowledgments

Sergio A Velastin has received funding from the Universidad Carlos III de Madrid, the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 600371, el Ministerio de Economía, Industria y Competitividad (COFUND2013-51509) el Ministerio de Educación, Cultura y Deporte (CEI-15-

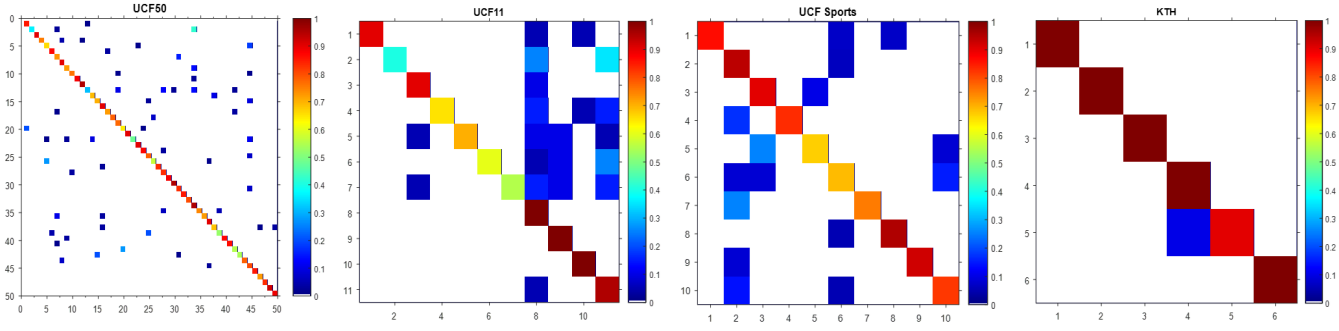

Fig. 10. Confusion matrices for the UCF50, UCF11, UCF-Sports and KTH datasets.

17) and Banco Santander. Authors also acknowledge support from the Directorate of ASR and TD, University of Engineering and Technology Taxila, Pakistan. .

References

Abdulmunem, A., Lai, Y., Sun, X., 2016. Saliency guided local and global descriptors for effective action recognition. Computational Visual Media 2, 97–106.

Dollár, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features, in: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, IEEE. pp. 65–72.

Duan, L., Xu, D., Tsang, I.W.H., Luo, J., 2012. Visual event recognition in videos by learning from web data. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 1667–1680.

Duta, I.C., Uijlings, J.R., Ionescu, B., Aizawa, K., Hauptmann, A.G., Sebe, N., 2017. Efficient human action recognition using histograms of motion gradients and vlad with descriptor shape information. Multimedia Tools and Applications , 1–28.

Gilbert, A., Illingworth, J., Bowden, R., 2011. Action recognition using mined hierarchical compound features. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 883–897.

Jain, A.K., 2010. Data clustering: 50 years beyond k-means. Pattern recognition letters 31, 651–666.

Klaser, A., Marszałek, M., Schmid, C., 2008. A spatio-temporal descriptor based on 3d-gradients, in: BMVC 2008-19th British Machine Vision Conference, British Machine Vision Association. pp. 275–1.

Kovashka, A., Grauman, K., 2010. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE. pp. 2046–2053.

Laptev, I., Lindeberg, T., 2003. Space-time interest points, in: 9th International Conference on Computer Vision, Nice, France, IEEE conference proceedings. pp. 432–439.

Liu, J., Luo, J., Shah, M., 2009. Recognizing realistic actions from videos in the wild, in: Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on, IEEE. pp. 1996–2003.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision 60, 91–110.

Mota, V.F., Souza, J.I., Araújo, A.d.A., Vieira, M.B., 2013. Combining orientation tensors for human action recognition, in: Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI-Conference on, IEEE. pp. 328–333.

Mukherjee, S., Biswas, S.K., Mukherjee, D.P., 2011. Recognizing human action at a distance in video by key poses. IEEE Transactions on Circuits and Systems for Video Technology 21, 1228–1241.

Nagel, H.H., Gehrke, A., 1998. Spatiotemporally adaptive estimation and segmentation of of-fields, in: European Conference on Computer Vision, Springer. pp. 86–102.

Niebles, J.C., Wang, H., Fei-Fei, L., 2008. Unsupervised learning of human action categories using spatial-temporal words. International journal of computer vision 79, 299–318.

Peng, X., Schmid, C., 2016. Multi-region two-stream r-cnn for action detection, in: European Conference on Computer Vision, Springer. pp. 744–759.

Peng, X., Wang, L., Wang, X., Qiao, Y., 2016. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding 150, 109–125.

Reddy, K.K., Shah, M., 2013. Recognizing 50 human action categories of web videos. Machine Vision and Applications 24, 971–981.

Rodriguez, M.D., Ahmed, J., Shah, M., 2008. Action mach a spatio-temporal maximum average correlation height filter for action recognition, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE. pp. 1–8.

Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: A local svm approach, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, IEEE. pp. 32–36.

Scovanner, P., Ali, S., Shah, M., 2007. A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th ACM international conference on Multimedia, ACM. pp. 357–360.

Wang, H., Kläser, A., Schmid, C., Liu, C.L., 2011. Action recognition by dense trajectories, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE. pp. 3169–3176.

Wang, H., Kläser, A., Schmid, C., Liu, C.L., 2013. Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision 103, 60–79.

Wang, H., Oneata, D., Verbeek, J., Schmid, C., 2016. A robust and efficient video representation for action recognition. International Journal of Computer Vision 119, 219–238.

Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C., 2009. Evaluation of local spatio-temporal features for action recognition, in: BMVC 2009-British Machine Vision Conference, BMVA Press. pp. 124–1.

Weinland, D., Ronfard, R., Boyer, E., 2011. A survey of vision-based methods for action representation, segmentation and recognition. Computer vision and image understanding 115, 224–241.

Willems, G., Tuytelaars, T., Van Gool, L., 2008. An efficient dense and scale-invariant spatio-temporal interest point detector. Computer Vision–ECCV 2008 , 650–663.

Wu, X., Xu, D., Duan, L., Luo, J., 2011. Action recognition using context and appearance distribution features, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE. pp. 489–496.

Yadav, G.K., Shukla, P., Sethfi, A., 2016. Action recognition using interest points capturing differential motion information, in: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE. pp. 1881–1885.

Yang, J., Ma, Z., Xie, M., 2015. Action recognition based on multi-scale oriented neighborhood features. International Journal of Signal Processing, Image Processing and Pattern Recognition 8, 241–254.

Zelnik-Manor, L., Irani, M., 2001. Event-based video analysis .